Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Impedance Matching of Humans ⇔ Machines in High-Q Information Retrieval Systems

Robert S. Bauer, Dan Brassil, Christopher Hogan, and Gina Taranto
H5
San Francisco, USA
{RBauer, DBrassil, CHogan, GTaranto}@H5.com

John Seely Brown
Deloitte Center for the Edge - San Jose, USA
&
University of Southern California – Los Angeles, USA
jsb@johnseelybrown.com

***Abstract*—Treating the information retrieval (IR) task as one of classification has been shown to be the most effective way to achieve high performance. In real-world Systems, a human is the ultimate determinant of relevance and must be integrated symbiotically into the control structures. We report on a hybrid, Human-Assisted Computer Classification system that opportunistically pairs processes of Active Learning and User Modeling to produce a high-Q computational engine. Top-down human goals are impedance-matched with bottom-up corpus analysis utilizing critical control loops. The System contributions of humans and machines as 'Proxy,' 'Assessor,' and 'Classifier' elements are blended through inter-related 'Model,' 'Match,' and 'Measure' processes (M^3) to achieve consistently high precision IR with high recall. We report results for over a dozen topics, with confirmation of internal measures from topic 103 of the 2008 TREC legal track's interactive task.**

***Keywords*—active learning, cybernetics for informatics, expert & knowledge-based systems, high-Q systems, human-machine cooperation & systems, impedance matching, information retrieval, knowledge acquisition in intelligent systems, knowledge engineering, knowledge representation, machine learning, personalization and user modeling, symbiotic theory formation**

I. INTRODUCTION

The challenge of any information retrieval (IR) system is to return a result that meets the goals of the User. In the ubiquitous paradigm of what is now regarded as 'Google search,' the IR task is one of providing the User with a few highly precise results. In such cases, the User goal may be to inform shopping, traveling, referencing, and other tasks where a handful of relevant documents is sufficient to achieve high system performance. A distinctly different challenge is when the User desires retrieval of *all* relevant documents and only documents that are relevant. Such cases are dominant in situations such as regulatory compliance, litigation, records retention, and business mergers/acquisitions. In this paper, we address the challenge of developing a System where retrieving *everything* of relevance is required by the User. The classic measure of performance for such a system is the requirement for high Recall (R) with high Precision (P) (herein referred to as "high-R&P") [1].

When the User requires *all* documents that are relevant to her need, the well-known system trade-off between P and R [2] produces an inaccurate result. Common search tools of all kinds either do not retrieve the bulk of relevant information (i.e., low R) or retrieve such a large number of irrelevant documents that relevant ones are only a small fraction of the output (i.e., low P: low signal-to-noise.) What is required is a System with high selectivity that can exhaustively identify documents meeting the User's goals. Such an automated system can only be developed by capturing and controlling the critical information flows among the human and computational agents. 'Consumer search' technologies and processes, no matter how 'advanced,' do not and cannot produce such a high-Q System.

Treating the IR task as one of classification has been shown to be the most effective way to achieve high performance [3]. All efforts to date to leverage these computational methods have assumed that the User will make final assessments of relevance from the output of the classification algorithm. Even if users could make perfect judgments, in most situations the sheer document volume makes it infeasible technically and economically to achieve high-P with high-R output. Thus the focus for a classification system must be to capture User intent effectively so that no further human intervention is necessary to achieve highly accurate document assessments. The challenge is one of impedance-matching human intentionality with algorithmic performance. By identifying actors (wo/man & machine,) enabling information flows, and controlling iterative processes, we achieve the first replicable, scalable IR System for high-P with high-R.

II. INFORMATION FLOWS

An IR System is one that produces an output that achieves a User's information goals when confronted with an input corpus. Given the resource constraints, knowledge transfer challenges, and human inconsistency, a high-R&P System cannot rely on human assessment as the final actor producing information output. Accuracy can only be achieved if User Goals are effectively represented in the control structures employed to train a highly selective computational classifier. We find that the most effective and efficient way to meet this requirement is to separate the information flows among 3 System elements: Proxy, Assessor, and Classifier. The functions of the first two are accomplished by humans while

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

the third element must be computational in order to achieve consistency during the course of exhaustive relevance identification. System elements, information flows, and control processes are depicted in Fig. 1.

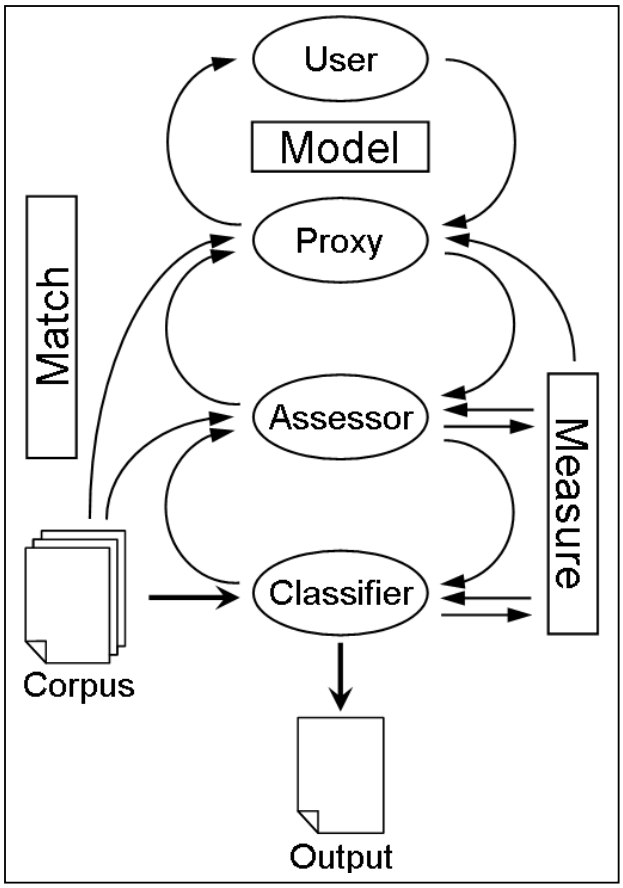

Figure 1. The M^3 processes(rectangles) and information flows (arrows) of a System that achieves a User's retrival objectives by providing relevant Output from a document Corpus with high Recall *and* high Precision.

A. User

The information goal of the User is always critical to successful IR. In 'consumer search,' where extremely high P with negligible R is the objective, the system is optimized to present candidate results to the User so she can make the final output selection because her goal is typically articulated through a single brief query of only a few words. Such a system does not attempt to model a User's intent; rather it relies on accumulated human judgment reflected in selections from innumerable like-queries (e.g., click-throughs) and other post-facto user indications such as PageRank™ analysis of webpage hyperlink structures [4]. High R&P tasks require detailed specification and iterative testing of user goals if an automated Classifier is to be trained for acceptable accuracy. Users are almost never IR experts and the state of technology is such that systems cannot adequately capture intent and iteratively improve performance in an efficient or effective manner with Users. We employ two separate roles of Proxy and Assessor to accomplish this impedance-Match of User goals with Classifier feature definition.

B. Proxy

User information during System development flows through a Proxy. This person is responsible for Modeling User goals and serving as User surrogate during development. Among other tasks, the Proxy brings expertise in questionnaire development, interviewing skills, experience with measurement protocols, and the behavior of the classification technology required to perform the Modeling process. The Proxy must reify the explicit User objective in much the same way that legal coding manuals are created for document reviewers [5]. However, the Modeling can be much more extensive when machines identify relevant documents rather than humans. The System benefits greatly from the Proxy eliciting tacit [6] 'know-how' from the User in addition to the traditional 'know-what' [7]. Indeed expert Proxies are individuals who use ambiguous exemplar documents to elucidate critical distinctions of relevance for the Classifier to be trained to identify.

C. Assessor

The Proxy does not work directly with the Classifier or document Corpus. This role is assigned to an Assessor who conducts detailed Corpus analysis, Classifier tuning, and measurement of System performance. One could say that the traditional knowledge engineer, who builds an expert systems, combine the roles of both Proxy and Assessor. Separating these information flows enables exploitation of distinctly different control loops during development. While the Proxy serves as surrogate for the User, the Assessor conducts the Measurement process and represents the machine (i.e., classification engine) in the processes discussed in the next section.

D. Classifier

For the purposes of this paper, the information flow, performance measurement, and evaluation of Classifier accuracy is of primary concern rather than technological details of the software employed. From a control-system perspective, computational Classifiers can be engineered to be considerable more comprehensive to discriminating document features than humans simply because the massive numbers of features that can be evaluated by machines. While any individual document characteristic can be analyzed more acutely by humans, the inherent limitation of our brains to sifting for 10 or fewer elements with 5 or less states disadvantages us from effective evaluation of the corpus even if we could be totally consistent among documents and with others assessments. We have found that the detailed Classifier methodology is less critical to accuracy than properly measuring Classifier performance and adjusting the control parameters to improve impedance Matching of the User Model to the Classifier.

III. CONTROL PROCESSES

Development of a high-Q System that encompasses the ensemble of information flows in Fig. 1 is equivalent to human-mediated symbiotic theory formation [8, 9] that encodes User IR goals into selective Classification features. Indeed relevance discrimination can be successfully modeled as

"iterative learning-loop complexes" [10]. The three inter-related processes that successfully leverage information flows in high-R&P Systems are referred to as M^3: Model, Match, and Measure.

A. Model

Relevance criteria must be established as a critical initial step in Modeling User goals. As noted, the Proxy's primary responsibility is to work with the User and the Assessor to produce a Model that can be richly represented by the Classifier. Detailed aspects of this process are discussed in other work [11].

B. Match

The Assessor's primary responsibility is to impedance Match the User Model developed by the Proxy with the Corpus so that the Classifier provides high-Q IR performance. Such a high-R&P System relies on the development of a symbiotic relationship between the top-down objectives of the User with the bottom-up characteristics of the Corpus. Such a process successfully explains over a decade of research into the sensemaking practices of intelligence analysts [12]. Crucial elements of process include detailed Corpus sampling and analysis, feature extraction, and sample testing.

C. Measure

Statistically valid measurements are required to achieve the impedance Matching discussed above. Therefore, it is the Assessor's responsibility to provide primary guidance to the Measurement process. Beyond providing proof of overall systems performance, measurements can be used to characterize the typicality of documents used as exemplars for User feedback. Of course measurements are also central to insuring that samples are representative of the Corpus and that relevance consistency is achieved. Without in-process measures, the validity of the Output meeting User objectives is pure speculation. While this may seem obvious, it is surprising that commercial IR 'accuracy' claims are often made without any valid, quantitative measures and that User's do not seek to understand estimated recall for highly precise Outputs.

IV. Results

Independent studies, where IR assessment protocols have been established by community consensus, are the best indicators of system effectiveness. The most comprehensive effort for high-R&P systems is the interactive task of NIST's Text Retrieval Conference (TREC) Legal Track [13].

A. System Performance

The results for one of the 3 tasks (T103) of the 2008 evaluation is reproduced in Fig.2 [14]. While only four teams participated in this inaugural effort, this year's participation involves many more institutions, both academic and commercial. The salient feature of these results is that a proof point now exists that high recall with high precision is possible. All prior TREC tracks have been strictly precision oriented, so it is necessary to establish foundational collections, tasks, and evaluations for this high-R&P systems.

All systems achieve high P, but the only system that also attains high R is the one that assessed relevance in a fully automated approach with impedance matching of user goals rather than human assessment of the final results.

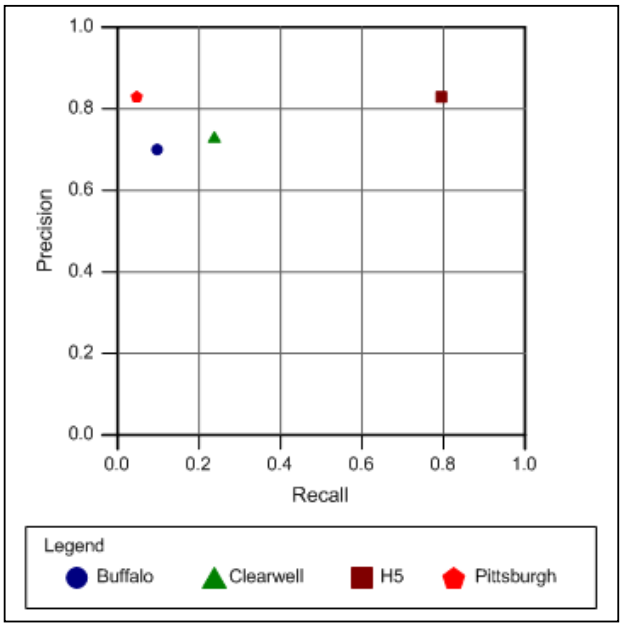

Figure 2. Recall and Precision of post-adjudicated results for high-OCR documents relevant to topic T103 of the interactive task of the 2008 TREC Legal track. Only the 'H5' System employed the impedance-matched, high-Q architecture in Fig. 1. This chart is reproduced from Fig. 3 of Ref. [14].

Two other independent studies provide important confirmation of System performance in this evaluation of approaches to high-R&P IR. Firstly, the most extensive controlled study of User classification efficacy is the 1985 research of Blair and Maron [15]. They instructed attorneys to review a corpus of just under 40,000 documents (consisting of about 350,000 pages) until they were confident that they had retrieved 75% of the relevant documents. While 79% of the documents were determined by independent measurement to be relevant, the text-retrieval 'system' that used computation to assist users actually retrieved less than 20% of the documents relevant to a particular search. This result is fully in line with the 3 data points in Fig. 2 indicating that not much has changed in computer-assisted human assessment performance even though the machine's technology has certainly employed algorithms and methods of much greater sophistication and complexity than used nearly a quarter century ago.

Kershaw conducted an independent study in 2005 of the human-assisted computer assessment System depicted in Fig. 1 [16]. 48,000 documents (230,000 pages) were assessed for relevance to three different topics by the machine classifier trained according to the M^3 processes discussed above. On average a cross the three topics, the impedance-matched high-Q System retrieved more than 95% of the relevant documents correctly with a precision of about 82%. This is in line with the TREC result for corresponding approach labeled 'H5' in Fig. 2.

It is important to establish that internal relevance measures employed in control loop development are valid indicators of

System performance. Table I. shows that our internal measurements are identical to TREC's independent evaluation within the limits of the Confidence Intervals (CI) [17].

TABLE I. FINAL RESULTS
FOR TOPIC 103 OF THE INTERACTIVE TASK OF THE 2008 TREC LEGAL TRACK. (THESE DATA DIFFER FROM THE TREC FINAL EVALUATION IN FIG. 2 BECAUSE THEY ARE BOTH PRE-ADJUDICATED AND NOT CORRECTED FOR OCR ERRORS.)

	Recall Est.	CI	Precision Est.	CI
TREC[a]	0.624	(0.579,0.668)	0.810	(0.795,0.824)
Internal[b]	0.687	(0.645,0.730)	0.823	(0.787,0.860)

a. 95% Confidence Interval (CI) b. 90% CI

(This Table is reproduced from Table 3 of Ref. [17].)

B. System Development

Employing M^3 processes to information flows in an impedance-matched high-Q System improves performance, enables responsiveness to changing requirements during development, and reduces uncertainty of the relevance assessments. Fig. 3 provides the evolution of performance measures during development of the 'H5' System for the TREC evaluation.

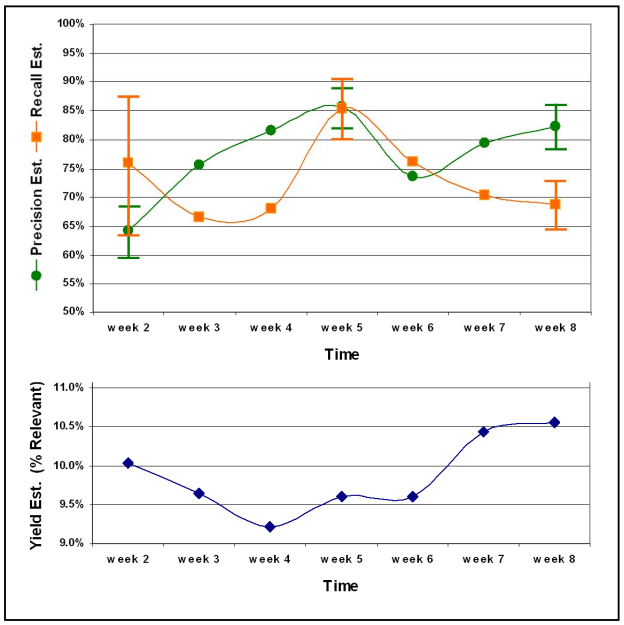

Figure 3. Internal measures of Precision, Recall, and Yield during System development of the 'H5' System for topic T103 of the interactive task of the 2008 TREC Legal track. Note that these data differ from the final result in Fig. 2 because thay are both pre-adjudicated and not corrected for OCR errors.

P, R, and Yield improve during System development, with tuning of each characteristics possible even though there are clear parametric interactions. As P and R were improved from weeks 2 through 5, the yield of relevant documents was reduced. Before the next iterative cycle of Match and Measure, the User (called the 'Topic Authority' by TREC) made changes to the assessments of relevant documents based on reconsideration of ambiguous exemplars that were presented during the Top-Down ⇔ Bottom-Up impedance Matching. This is precisely the iterative behavior seen in multiple sense-making loops employed by intelligence analysts [12]. It indicates the importance of understanding the information flows and control loops that invariably must be adjusted during the course of development. In this case, the Proxy had to alter the User Model after week 5 thereby requiring the Assessor to re-train the classifier. The Assessor's performance Measure and resulting feature adjustment for the new characteristics of the User Model enabled recovery of high selectivity. With a TREC established deadline at week 8, the final iterations focused on yield improvement while the accumulated sampling statistics led to reduced uncertainty for both P and especially R estimates.

Initial automated Classifier performance is often no better than for approaches (including final human-assessment) that do not impedance Match top-down User Models with a plethora of bottom-up Corpus features. However, the type of controlled performance improvement demonstrated for TREC topic 103 (Fig. 3) is achievable consistently as shown in Fig. 4 for 12 topics to which M^3 processes were employed [18].

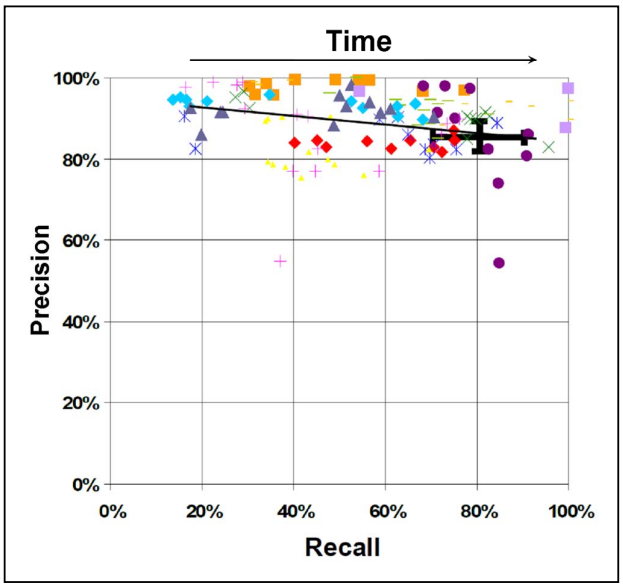

Figure 4. Sampled Corpus tests during high-R&P System developoment for 12 topics. Each topic is a different symbol (and color.) Retrieval performance for all topics moves through time from left to right. This chart is reproduced from Fig. 3 of Ref. [18].

The temporal mean for the 12 topics in Fig. 4 (i.e., black line from P~93% @ R~18% to P~85% @ R~80%) is typical for Classifier development that employs the impedance-matching, high-Q architecture in Fig. 1. Comparison between these evolutions with that depicted in Fig. 3 demonstrates a consistent ability to achieve controllably high-R&P performance.

V. CONCLUSION

We report the achievement of a replicable, scalable high-R&P IR System by capturing and controlling necessary information flows between human and machine elements. The prevailing practice for attempting to achieve high-R&P Systems uses humans as the final Assessors of Classifier results to produce IR Output. Such an approach has *never* been demonstrated to work with any rigorous testing protocol. This should not be surprising considering the inaccuracies of transferring goals of the User with enough specificity and consistency to surrogates. Indeed practitioners widely report inconsistency in assessments among surrogates and drifts in quality over the course of an assessment period. Such approaches are inherently incapable of adjusting to evolution in the User's explicit understanding of her goal during the course of an IR task.

Contrary to such prevailing approaches, we have shown that Proxies and Assessors can leverage information loops to adapt an automated Classifier to evolutionary IR realities. Proxies are necessary to accurately Model User goals and interact with Assessors, who impedance-Match the User's IR requirements to Classifier technologies. Rigorous Measurement and iterative, active learning loops provide sufficient control for achieving High-Q, selective output of relevant documents from a Corpus.

Perhaps future ontological-based platforms, with intuitive, interactive user interfaces, will be developed wherein users can efficiently and effectively input their goals, and iterative assessment with course-correction will be mechanistically performed to train a computer classifier. In the meantime, this paper outlines an independently validated approach to developing an IR System for high P with high R. The high-performance results reported for 16 topics in three corpora employed three agents (2 wo/man & 1 machine,) who participated in three iterative, inter-related M^3 processes. This architecture enables adaptation to and control of information flows that produces an Output that achieves a User's retrieval goals when confronted with an input Corpus.

REFERENCES

[1] R. Baeza-Yates, B. Ribeiro-Neto (1999). Modern Information Retrieval. New York: ACM Press, Addison-Wesley. p. 75 ff.

[2] C. Buckley and E. M. Voorhees, Chapter 3, "Retrieval System Evaluation" in Voorhees & Harman's "TREC: Experiment and Evaluation in Information Retrieval" (2005) MIT Press, p. 62, Fig. 3.

[3] C. D. Manning, P. Raghavan and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press. 2008

[4] L. Page, S. Brin, R. Motwani, T. Winograd, (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

[5] P. B. Madden, "Information Management in Complex Litigation," in Litigation 4 (1977), p. 12.

[6] M. Polanyi, The Tacit Dimension (Garden City, NY: Doubleday, 1966)

[7] J. S. Brown, "Learning in the Digital Age," in The Internet & the University: Forum 2001 edited by Maureen Devlin, Richard Larson and Joel Meyerson, pp. 65-91. Published as a joint project of the Forum for the Future of Higher Education and EDUCAUSE, 2002.

[8] J. S. Brown. A Symbiotic Theory Formation System. PhD thesis, University of Michigan, 1972.

[9] J. Roschelle and S. Teasley. The construction of shared knowledge in collaborative problem solving. In C. O'Malley, editor, Computer-supported collaborative learning, pages 69–77. Springer-Verlag, Heidelberg, Germany, 1995.

[10] D. M. Russell, M. J. Stefik, P. L. Pirolli, S. K. Card, "The Cost Structure of Sensemaking," Proceeding of InterCHI, ACM. (1993), 269-276.

[11] D. Brassil, C. Hogan, S. Attfield, "The Centrality of User Modeling to High Recall with High Precision Search," IEEE - SMC 2009 Proceedings.

[12] L. Takayama, S. K. Card, "Tracing the Microstructure of Sensemaking," CHI workshop on Sensemaking, Florence, Italy (April 6, 2008) Fig. 2.

[13] National Institute of Standards and Technology (NIST) - TREC 2008 Legal Track: Interactive Task guidelines.

[14] D. Oard, B. Hedin, S. Tomlinson, and J. Baron. Overview of the TREC 2008 legal track. In Proceedings of The Seventeenth Text Retrieval Conference (TREC-2008), 2008.

[15] D. C. Blair, M. E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," Communications of the ACM, 28, #3 (1985), 289-299.

[16] A. Kershaw. "Automated Document Review Proves its Reliability." Digital Discovery & e-Evidence, (November, 2005.)

[17] C. Hogan, D. Brassil, S. Rugani, J. Reinhart, M. Gerber, T. Jade, "H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement," NIST TREC Conference (November, 2008,) Table 3.

[18] R. S. Bauer, T. Jade, B. Hedin, and C. Hogan, "Automated Legal Sensemaking: The Centrality of Relevance and Intentionality," DESI II Workshop, London (June 25, 2008.). 301, 1982].